Biometrika (2012), **xx**, x, *pp*. 1–13
© 2007 Biometrika Trust
Printed in Great Britain

arXiv:1202.2683v1 [stat.ME] 13 Feb 2012

Retrospective–prospective symmetry for the Bayesian analysis of case-control studies

BY SIMON P. J. BYRNE
Department of Statistical Science, University College, London WC1E 6BT, U.K.
byrne@stats.ucl.ac.uk

AND A. PHILIP DAWID
Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, U.K.
a.p.dawid@statslab.cam.ac.uk

SUMMARY

The seminal paper of Prentice & Pyke (1979) established that the maximum likelihood estimator for the odds-ratio of a case-control study is that of a logistic regression. In other words, the incorrect prospective model is equivalent to the correct retrospective model. We identify necessary and sufficient conditions for the corresponding result in a Bayesian analysis, that is, that the posterior distribution for the odds-ratio be the same under both the prospective and retrospective likelihoods. These conditions can be used to derive a parametric family of prior laws that can be used for such an analysis.

Some key words: Case-control study; logistic regression; retrospective likelihood; hyper Markov law; conditional independence.

In order to estimate the risk factors for a disease (or any other binary outcome), there are two basic approaches: a *prospective* or *cohort* study, in which subjects are selected from the population, possibly based on their risk factors, and observed to determine if the disease arises; and a *case-control* or *retrospective* study, in which random samples are taken from both the population with the disease (cases), and the population without (controls), and the relative frequencies of the risk factors in the two samples is then recorded.

Let Y be the outcome variable taking values in $\{0,1\}$, corresponding to the absence or presence of disease, respectively. Let X be the vector of covariates (risk factors) taking values in $\mathcal{X} \subseteq \mathbb{R}^k$. In a prospective study we are sampling from the conditional distribution of Y given X. Under a proportional odds assumption, the model is that of a logistic regression,

$$p(y \mid x, \alpha, \beta) = \frac{e^{y(\alpha+\beta^\top x)}}{1+e^{\alpha+\beta^\top x}}, \quad \alpha \in \mathbb{R}, \beta \in \mathbb{R}^k. \tag{1}$$

On the other hand, a case-control study will result in observations from the conditional distribution of X given Y. In this case, specifying a probabilistic model becomes much more difficult, particularly if $\mathcal{X}$ is infinite.

Despite these difficulties, case-control studies are often desirable, or in some cases unavoidable, particularly where the disease is relatively rare or the time until diagnosis is long, as the costs of obtaining a sufficient sample size for a prospective study are likely to be prohibitive.

Prentice & Pyke (1979) showed that the maximum likelihood estimator of the log-odds ratio parameter β and its asymptotic covariance, could simply be found by a logistic regression. In other words, we can use the prospective model to analyse data gathered retrospectively. This particular result has been widely applied in epidemiology and other areas.

In this paper, we identify the analogous result for the Bayesian case: that is, the conditions under which the posterior distribution for β can be computed using the prospective likelihood instead of the retrospective.

The simplest model of a single binary covariate, where $\mathcal{X} = \{0, 1\}$, has been well explored in literature: Zelen & Parker (1986), Nurminen & Mutanen (1987), Marshall (1988) and Ashby et al. (1993) have all characterized such an analysis, which consists of computing the posterior distribution of the log odds ratio of a 2×2 contingency table under a Dirichlet prior. In the case where the covariates are categorical, that is where $\mathcal{X}$ is finite, Seaman & Richardson (2004) identified a class of improper priors that satisfy the desired properties. This class was further expanded by Staicu (2010).

We show that the basis of this prospective–retrospective symmetry is due to "independence" of the parameters: the original result of Prentice & Pyke (1979) can be explained through the variation independence in the parameter space, and that the corresponding Bayesian result will occur when the prior law exhibits analogous probabilistic independence. Furthermore, we arrive at the same class of prior laws as Staicu (2010) via a different route, and demonstrate how they can be extended to stratified designs.

However this is not the only approach for Bayesian analysis of case-control data. With the advent of computational tools such as MCMC, the retrospective likelihood need not present such an obstacle. Indeed this path has been well followed in the literature, as reviewed in Mukherjee et al. (2005). For example, Müller & Roeder (1997), Seaman & Richardson (2001) and Gustafson et al. (2002) have pursued this approach. In particular, Gustafson et al. (2002) note that in general the prospective posterior can serve as a useful approximation to the retrospective posterior, and use this as the basis of an importance sampling scheme.

1. NOTATION AND DEFINITIONS

Throughout the paper, (X, Y) will denote a single joint observation from the specified model, and $(X^{(n)}, Y^{(n)})$ a sequence of n such observations; p will be denote density of the model (with respect to the appropriate measure), with variables indicating the context.

We recall the notation and definitions from Dawid & Lauritzen (1993). If θ denotes a joint probability distribution for (X, Y), then θ_X and θ_Y will denote the corresponding marginal distributions of X and Y respectively. Furthermore, $\theta_{Y|X=x}$ will be the conditional distribution of Y given $X = x$, and $\theta_{Y|X} = \{\theta_{Y|X=x} : x \in \mathcal{X}\}$ will be the family of all such conditional distributions, and likewise for $\theta_{X|Y}$.

A *model* is a set Θ of such joint probability distributions θ. For any two functions ϕ, τ on Θ, we define the *conditional range* of ϕ given $\tau = t$ to be $\{\phi(\theta) : \theta \in \Theta \text{ and } \tau(\theta) = t\}$. Furthermore, ϕ is said to be *variation independent* of τ, written $\phi \ddagger \tau$, if this is constant for all values of t; in other words, if (ϕ, τ) takes values in a product space. In a similar manner, we can define *conditional variational independence* (see Dawid & Lauritzen, 1993).

A model is called *strong meta Markov* if

$$\theta_X \ddagger \theta_{Y|X} \qquad \text{and} \qquad \theta_Y \ddagger \theta_{X|Y}. \tag{2}$$

We define a *law* $\pounds$ to be a probability distribution over a model. We say that a law is *strong hyper Markov* if we replace the variation independence of (2) with probabilistic independence

(denoted by $\perp\!\!\!\perp$) under $\pounds$:

$$\theta_X \perp\!\!\!\perp \theta_{Y|X} \qquad \text{and} \qquad \theta_Y \perp\!\!\!\perp \theta_{X|Y} \quad [\pounds].$$

As variation independence is a necessary condition for probabilistic independence, a necessary (but not sufficient) condition for a law to be strong hyper Markov is that its support be a strong meta Markov model.

We use the relation $\phi \simeq \psi$ to denote the existence of a bijective function between ϕ and ψ. For example, we have $\theta \simeq (\theta_X, \theta_{Y|X}) \simeq (\theta_Y, \theta_{X|Y})$.

LEMMA 1. *For the above logistic model,*

$$\theta_{Y|X} \simeq (\alpha, \beta) \qquad \textit{and} \qquad \theta_{X|Y} \simeq (\theta_{X|Y=0}, \beta).$$

Proof. The first equivalence follow from (1), and the second from Bayes theorem:

$$\frac{\mathrm{d}\theta_{X|Y=1}}{\mathrm{d}\theta_{X|Y=0}}(x) = \frac{\theta_{Y|X=x}(1)}{\theta_{Y|X=x}(0)} \frac{\theta_Y(0)}{\theta_Y(1)} \propto e^{\beta^\top x}.$$

□

The usual definition of independence does not apply in the case where $\pounds$ is improper, so instead we define $\phi \perp\!\!\!\perp \tau$ to mean that the joint density factorizes into a function of ϕ and a function of τ. Owing to the problems of marginalising improper distributions (see Dawid et al., 1973), this only makes sense if $\theta \simeq (\phi, \tau)$.

2. MAXIMUM LIKELIHOOD ESTIMATORS

Prentice & Pyke (1979) showed that the maximum likelihood odds-ratio estimators obtained from a case-control study have the same values and asymptotic properties as those arising from a prospective study; in particular, they can be computed from a prospective logistic regression. This can be demonstrated using the strong meta Markov property.

LEMMA 2. *Let* $\Theta \simeq \Theta_X \times \Theta_{Y|X}$*, where* Θ_X *is the family of all probability distributions over* $\mathcal{X}$*, and* $\Theta_{Y|X}$ *is the family of all conditional distributions with densities of the form in* (1). *Then the corresponding family of joint distributions* Θ *is strong meta Markov, that is,*

$$\theta_X \ \ddagger\ (\alpha, \beta) \qquad \textit{and} \qquad \theta_Y \ \ddagger\ (\theta_{X|Y=0}, \beta).$$

Proof. These properties are essentially a reformulation of Müller & Roeder (1997, Lemmas 1 and 2). By definition $\theta_X \ \ddagger\ \theta_{Y|X}$. It remains to show variation independence in the opposite direction.

For any θ_X and $\theta_{Y|X}$, the joint distribution θ has a density of the form

$$p(x, y \mid \theta) = \frac{e^{y(\alpha + \beta^\top x)}}{1 + e^{\alpha + \beta^\top x}} p(x \mid \theta_X).$$

Therefore the marginal distribution θ_Y is Bernoulli, with parameter γ taking values on the interval $(0, 1)$, where

$$\gamma = p(y = 1 \mid \theta_Y) = \int_{\mathcal{X}} \frac{e^{\alpha + \beta^\top x}}{1 + e^{\alpha + \beta^\top x}} p(x \mid \theta_X)\, dx \tag{3}$$

and the conditional distribution of X given Y has density of the form

$$p(x \mid y, \theta_{X|Y}) = \frac{p(x, y \mid \theta)}{\gamma^y (1 - \gamma)^{1-y}} = \frac{e^{y(\alpha - \log\frac{\gamma}{1-\gamma} + \beta^\top x)}}{(1 - \gamma)(1 + e^{\alpha + \beta^\top x})} p(x \mid \theta_X). \tag{4}$$

For any $\gamma' \in (0,1)$, define $\theta' \simeq (\theta'_X, \theta'_{Y|X})$, where

$$\theta'_{Y|X} \simeq (\alpha', \beta) \in \Theta_{Y|X} \quad \text{with} \quad \alpha' = \alpha - \log \frac{\gamma}{1-\gamma} + \log \frac{\gamma'}{1-\gamma'}, \tag{5}$$

and θ'_X has density

$$p(x \mid \theta'_X) = \frac{(1-\gamma')(1+e^{\alpha'+\beta^\top x})}{(1-\gamma)(1+e^{\alpha+\beta^\top x})} p(x \mid \theta_X).$$

By the definition of γ in (3), it can be shown that this integrates to 1, hence $\theta'_X \in \Theta_X$. Furthermore, by matching terms in (4), then $\theta_{X|Y} = \theta'_{X|Y}$. Since $\theta'_Y \simeq \gamma'$ can be chosen arbitrarily, it follows that $\theta_Y \ddagger \theta_{Y|X}$. □

We can use the fact that variation independence satisfies the same properties as conditional independence (Dawid & Lauritzen, 1993):

COROLLARY 1. *Under the joint logistic model of Lemma 2,*

$$\theta_X \ddagger \alpha \mid \beta \qquad \textit{and} \qquad \theta_Y \ddagger \theta_{X|Y} \mid \beta.$$

The logistic model has other variation independence properties:

COROLLARY 2. *Under the joint logistic model of Lemma 2,*

$$(\theta_X, \theta_Y) \ddagger \beta.$$

Proof. We have $\theta_X \ddagger (\alpha, \beta)$, and for any θ_Y, we can choose α' as in (5). □

THEOREM 1. *Suppose we have a joint model as in Lemma 2. Then the profile likelihood function for the odds ratio β is the same for both the retrospective model $\Theta_{X|Y}$ and the prospective model $\Theta_{Y|X}$, up to proportionality.*

Proof. This proof follows a similar argument as Dawid & Lauritzen (1993, Lemma 4.10). The joint density for the model θ can be written as

$$p(x, y \mid \theta) = p(x|\theta_X)p(y \mid x, \alpha, \beta) = p(y \mid \theta_Y)p(x \mid y, \theta_{X|Y=0}, \beta).$$

Therefore the profile likelihood for the joint model can be written in terms of the prospective model:

$$L_{\mathrm{p}}^{\text{joint}}(\beta) = \max_{\alpha, \theta_X} p(x \mid \theta_X) p(y \mid x, \theta_{Y|X}). \tag{6}$$

By the conditional variation independence α and θ_X given β of Corollary 1, the factors of (6) can be profiled separately, so that

$$L_{\mathrm{p}}^{\text{joint}}(\beta) \propto \max_{\alpha} p(y \mid x, \alpha, \beta) = L_{\mathrm{p}}^{\text{pro}}(\beta),$$

where $L_{\mathrm{p}}^{\text{pro}}$ denotes the profile likelihood of the prospective model. The same argument applies to the retrospective profile likelihood $L_{\mathrm{p}}^{\text{ret}}(\beta)$:

$$L_{\mathrm{p}}^{\text{joint}}(\beta) \propto \max_{\theta_{X|Y=0}} p(x \mid y, \theta_{X|Y=0}, \beta) = L_{\mathrm{p}}^{\text{ret}}(\beta).$$

□

From this we obtain the result of Prentice & Pyke (1979):

COROLLARY 3. *For data observed in a case-control study, the maximum likelihood estimator of the log odds parameter $\hat{\beta}$ and its asymptotic covariance can be computed as if the data were observed prospectively, that is, using logistic regression.*

Proof. The maximum likelihood estimator is a function of the profile likelihood, as is its asymptotic covariance (see Patefield, 1985). □

The same argument can also be applied to the value, but not the covariance, of any penalized logistic regression estimator of the form

$$\arg\max_{\alpha,\beta} \left\{ \log p(y \mid x, \alpha, \beta) + \phi(\beta) \right\}.$$

Examples of such estimators include ridge regression, where $\phi(\beta) \propto \|\beta\|_2$, and lasso, where $\phi(\beta) \propto \|\beta\|_1$. Such methods have proven successful in genome-wide association studies, which involve case-control data with extremely high-dimensional covariates (Park & Hastie, 2008; Wu et al., 2009).

3. BAYESIAN ANALYSIS OF CASE-CONTROL STUDIES

We now investigate how these results correspond to a Bayesian analysis. We use π to denote the density of the prior law, and π^{pro} and π^{ret} to denote the densities of the posterior laws $\mathcal{L}^{\text{pro}}$ and $\mathcal{L}^{\text{ret}}$ under prospective and retrospective likelihoods, respectively:

$$\pi^{\text{pro}}(\alpha, \beta \mid x^{(n)}, y^{(n)}) \propto \pi(\alpha, \beta)\, p(y^{(n)} \mid x^{(n)}, \alpha, \beta)$$
$$\pi^{\text{ret}}(\theta_{X|Y=0}, \beta \mid x^{(n)}, y^{(n)}) \propto \pi(\theta_{X|Y=0}, \beta)\, p(x^{(n)} \mid y^{(n)}, \theta_{X|Y=0}, \beta)$$

Furthermore, we will use $\bar{p}$ to denote the density of the marginal model, where parameters have been integrated out (using the prior law), for example

$$\bar{p}(y^{(n)} \mid x^{(n)}, \beta) = \int p(y^{(n)} \mid x^{(n)}, \alpha, \beta) \pi(\alpha \mid \beta)\, d\alpha.$$

In other words, when interpreted as a function of β, $\bar{p}(y^{(n)}|x^{(n)}, \beta)$ is the marginal likelihood for β.

We now present the key result of this section.

THEOREM 2. *Let $\mathcal{L}(\tilde{\theta})$ be a prior law for the joint parameters of the logistic model. Then the posterior marginal law for $\tilde{\beta}$ is the same under both prospective and retrospective likelihood for all sample sizes n, and all possible observations $(x^{(n)}, y^{(n)})$, if and only if*

$$\tilde{\beta} \perp\!\!\!\perp \tilde{\theta}_X \qquad \textit{and} \qquad \tilde{\beta} \perp\!\!\!\perp \tilde{\theta}_Y \qquad [\mathcal{L}]. \tag{7}$$

Proof. First, the marginal posterior densities for $\tilde{\beta}$ can be written as

$$\pi^{\text{pro}}(\beta \mid x^{(n)}, y^{(n)}) \propto \pi(\beta)\, \bar{p}(y^{(n)} \mid x^{(n)}, \beta)$$
$$\pi^{\text{ret}}(\beta \mid x^{(n)}, y^{(n)}) \propto \pi(\beta)\, \bar{p}(x^{(n)} \mid y^{(n)}, \beta),$$

where $\bar{p}$ denotes the marginal model. Hence the marginal posteriors are equal if and only if the retrospective and prospective marginal likelihoods for β are proportional, for $\pi(\beta) > 0$. In other words, whenever there exists a function k such that

$$\bar{p}(x^{(n)} \mid y^{(n)}, \beta) = \bar{p}(y^{(n)} \mid x^{(n)}, \beta)\; k(x^{(n)}, y^{(n)}). \tag{8}$$

These models are also related through the joint model

$$\bar{p}(x^{(n)} \mid y^{(n)}, \beta)\, \bar{p}(y^{(n)} \mid \beta) = \bar{p}(y^{(n)} \mid x^{(n)}, \beta)\, \bar{p}(x^{(n)} \mid \beta),$$

therefore (8) is equivalent to

$$\bar{p}(x^{(n)} \mid \beta) = \bar{p}(y^{(n)} \mid \beta)\, k(x^{(n)}, y^{(n)}). \tag{9}$$

Since $X^{(n)} \perp\!\!\!\perp \tilde{\beta} \mid \tilde{\theta}_X$, we can write the marginal model for $X^{(n)} \mid \tilde{\beta}$ as

$$\bar{p}(x^{(n)} \mid \beta) = \int_{\Theta_X} \left\{ \prod_{i=1}^{n} p(x_i \mid \theta_X) \right\} \pi(\theta_X \mid \beta)\, d\theta_X. \tag{10}$$

Therefore, if $\tilde{\theta}_X \perp\!\!\!\perp \tilde{\beta}$, then $\bar{p}(x^{(n)} \mid \beta)$ must be constant in β, and similarly for $\bar{p}(x^{(n)} \mid \beta)$ if $\tilde{\theta}_Y \perp\!\!\!\perp \tilde{\beta}$. Hence (7) implies (9).

To show the converse, suppose that (9) holds for all n and values of $(x^{(n)}, y^{(n)})$. As $\bar{p}(x^{(n)} \mid \beta)$ is a density, it must be proportional to $k(x^{(n)}, y_0^{(n)})$, for any fixed $y_0^{(n)}$, and so $X^{(n)}$ is independent of $\tilde{\beta}$.

Now $\bar{p}(x^{(n)} \mid \beta)$ is the density of a mixture of independent and identically distributed variables, and the mixing measure of such an infinite sequence is uniquely determined (Aldous, 1985, Lemma 2.15). It follows that $\pi(\theta_X \mid \beta)$ must be independent of β, and hence $\tilde{\theta}_X \perp\!\!\!\perp \tilde{\beta}$. The same argument holds for $\tilde{\theta}_Y$. □

Several authors have identified similar results. Notably, Müller & Roeder (1997) appear to have almost identified the conditions in (7), but then incorrectly claim that the "argument about the retrospective likelihood only carries over to posterior inference on β if α and β are independent and θ_X is not otherwise constrained". This misconception appears to be due to the fact that although there is a one-to-one mapping between α and θ_Y, this mapping is itself dependent on β, through (3). Unfortunately, this means that the Dirichlet process mixture they propose does not satisfy the required properties.

Example 1. Any law $\pounds(\tilde{\theta})$ with the property

$$(\tilde{\theta}_X, \tilde{\theta}_Y) \perp\!\!\!\perp \tilde{\beta} \qquad [\pounds].$$

We can construct such a law from two arbitrary laws $\pounds_{\mathrm{m}}(\tilde{\theta})$ and $\pounds_{\mathrm{o}}(\tilde{\theta})$, on taking $\pounds$ to be the product law of their projections $\pounds_{\mathrm{m}}(\tilde{\theta}_X, \tilde{\theta}_Y)$ and $\pounds_{\mathrm{o}}(\tilde{\beta})$. By Corollary 2, there will exist a $\tilde{\theta}$ with these marginals, and since

$$\tilde{\theta} \simeq \big(\tilde{\theta}_X, \alpha(\tilde{\theta}_X, \tilde{\theta}_Y, \tilde{\beta}), \tilde{\beta}\big) \simeq (\tilde{\theta}_X, \tilde{\theta}_Y, \tilde{\beta}),$$

such a law would be uniquely determined.

Unfortunately, such a law would probably not be all that useful, as it would still require computing the integral

$$\bar{p}(y \mid x, \beta) = \int_{\Theta_X \times \Theta_Y} \frac{e^{\alpha(\beta, \theta_X, \theta_Y) + \beta^\top x}}{1 + e^{\alpha(\beta, \theta_X, \theta_Y) + \beta^\top x}}\, d\pounds_{\mathrm{m}}(\theta_X, \theta_Y),$$

which may not be any easier than the retrospective likelihood.

In order to avoid the need to compute such integrals, we can require $\tilde{\alpha}$ and $\tilde{\theta}_X$ to be independent, such as in strong hyper Markov laws.

COROLLARY 4. *If $\mathcal{L}(\tilde{\theta})$ is strong hyper Markov, that is,*

$$(\tilde{\alpha}, \tilde{\beta}) \perp\!\!\!\perp \tilde{\theta}_X \qquad \textit{and} \qquad (\tilde{\theta}_{X|Y=0}, \tilde{\beta}) \perp\!\!\!\perp \tilde{\theta}_Y \qquad [\mathcal{L}],$$

then the posterior law for $\tilde{\beta}$ is the same under both the prospective and the retrospective likelihood.

For the case that $\mathcal{X}$ is finite, conditions equivalent to the strong hyper Markov property were shown to be sufficient in a 2007 University of Bristol technical report by A.-M. Staicu.

The problem of model comparison for case-control studies has received comparatively little attention in the literature, particularly for Bayesian analyses. However we can derive a result similar to that of Theorem 2.

THEOREM 3. *If $\mathcal{L}_1(\tilde{\theta})$ and $\mathcal{L}_2(\tilde{\theta})$ have the same marginal laws for $\tilde{\theta}_X$ and $\tilde{\theta}_Y$, then the Bayes factor between the prospective models is equal to the Bayes factor between the retrospective models.*

Proof. Let $\tilde{M}$ take values 1 and 2 each with probability $1/2$, and, given $\tilde{M} = j$, let the conditional law of $\tilde{\theta}$ be $\mathcal{L}_j$. In the resulting joint law $\mathcal{L}^*$ for $(\tilde{\theta}, \tilde{M})$, when the conditions of the theorem hold we shall have

$$\tilde{M} \perp\!\!\!\perp \tilde{\theta}_X \qquad \text{and} \qquad \tilde{M} \perp\!\!\!\perp \tilde{\theta}_Y \qquad [\mathcal{L}^*].$$

By the same argument as for Theorem 2, the posterior probabilities, and hence the Bayes factors, must be equal. □

4. STRONG HYPER MARKOV LAWS FOR LOGISTIC REGRESSION

Given the results of Corollary 4, we now investigate various strong hyper Markov laws for use as prior laws in case-control studies.

4·1. *A single binary covariate*

In the case of a single binary covariate, $\mathcal{X} = \{0, 1\}$, the logistic model is just a reparametrization of the 2×2 contingency table.

Example 2. The simplest strong hyper Markov law for this model is the Dirichlet law $\mathcal{L}(\tilde{\theta}) = \mathcal{D}(a_{xy})$, with density

$$\pi(\theta) = \frac{1}{B(\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11})} \theta_{00}^{a_{00}-1} \theta_{01}^{a_{01}-1} \theta_{10}^{a_{10}-1} \theta_{11}^{a_{11}-1},$$

where $\theta_{xy} = p(X = x, Y = y \mid \theta)$. This law has been well explored in the literature, in particular by Altham (1969), who investigated log odds ratio parameter: and was later used in the context of case-control studies by Zelen & Parker (1986), Nurminen & Mutanen (1987), Marshall (1988) and Ashby et al. (1993).

By reparametrizing $\theta_{xy} = \frac{e^{y(\alpha+\beta x)}}{1+e^{\alpha+\beta x}} {\theta_{0+}}^{1-x} {\theta_{1+}}^{x}$, we find $\mathcal{L}(\tilde{\theta}_{x+}) = \mathcal{B}(a_{0+}, a_{1+})$, and

$$\pi(\alpha, \beta) = \frac{e^{\alpha a_{01}} e^{(\alpha+\beta) a_{11}}}{(1 + e^{\alpha})^{a_{0+}} (1 + e^{\alpha+\beta})^{a_{1+}}}.$$

However the family of strong hyper Markov laws on 2×2 tables is more general than this. Geiger & Heckerman (1997, equation 10) note that a law with full support is strong hyper

Markov, which they term "global parameter independence", if and only if it has a density of the form

$$\pi(\theta) \propto h\left(\frac{\theta_{00}\theta_{11}}{\theta_{01}\theta_{10}}\right)\theta_{00}^{\alpha_{00}-1}\theta_{01}^{\alpha_{01}-1}\theta_{10}^{\alpha_{10}-1}\theta_{11}^{\alpha_{11}-1}, \tag{11}$$

for a positive Lebesgue integrable function h. The corresponding density of $\pounds(\tilde{\alpha}, \tilde{\beta})$ is

$$\pi(\alpha, \beta) = g(\beta)\frac{e^{\alpha a_{01}}e^{(\alpha+\beta)a_{11}}}{(1+e^{\alpha})^{a_{0+}}(1+e^{\alpha+\beta})^{a_{1+}}},$$

where $g(\beta) = h(e^{\beta})$.

4·2. *Finite covariate space*

A more general case is where $\mathcal{X}$ is larger but still finite, for example a model with multiple categorical covariates. Prior specification is now not so simple: the proportional odds constraint implies that the logistic model will be confined to a submanifold of the probability simplex of the full $|\mathcal{X}| \times 2$ contingency table.

We solve this problem by adapting the conditioning procedure of Dawid & Lauritzen (2001, section 4) for constructing laws on nested models, by firstly choosing an arbitrary strong hyper Markov law $\pounds'(\tilde{\theta})$ for the saturated model on $\mathcal{X} \times \{0, 1\}$, and then constructing the law $\pounds$ from $\pounds'$ conditional on $\tilde{\theta}$ satisfying the proportional odds requirement.

As Dawid & Lauritzen (2001) emphasized, the Borel–Kolmogorov paradox shows that there is no unique way to condition on a submodel. Furthermore, in selecting the method of conditioning, we need to ensure that it preserves the strong hyper Markov property.

We assume that there exists $x_1, \ldots, x_{k+1} \in \mathcal{X}$ such that $(1, x_1), (1, x_2), \ldots, (1, x_{k+1})$ are linearly independent, since otherwise β is not identifiable. We can reparametrize the saturated model as

$$p(y \mid x, \alpha, \beta, \eta) = \frac{e^{y(\alpha+\beta^{\top}x+\eta_x)}}{1+e^{\alpha+\beta^{\top}x+\eta_x}},$$

where $\eta_x = 0$ if $x = x_1, \ldots, x_{k+1}$. Then $\theta_{Y|X} \simeq (\alpha, \beta, \eta)$ and $\theta_{X|Y} \simeq (\theta_{X|Y=0}, \beta, \eta)$, and hence if $\pounds'$ is strong hyper Markov:

$$(\tilde{\alpha}, \tilde{\beta}, \tilde{\eta}) \perp\!\!\!\perp \tilde{\theta}_X \qquad \text{and} \qquad (\tilde{\theta}_{X|Y=0}, \tilde{\beta}, \tilde{\eta}) \perp\!\!\!\perp \tilde{\theta}_Y \qquad [\pounds'].$$

Note that the logistic model is the manifold defined by $\eta = 0$. Furthermore,

$$(\tilde{\alpha}, \tilde{\beta}) \perp\!\!\!\perp \tilde{\theta}_X \bigm| \tilde{\eta} \qquad \text{and} \qquad (\tilde{\theta}_{X|Y=0}, \tilde{\beta}) \perp\!\!\!\perp \tilde{\theta}_Y \bigm| \tilde{\eta} \qquad [\pounds'].$$

Hence $\pounds(\tilde{\theta})$ defined as $\pounds'(\tilde{\theta} \mid \tilde{\eta} = 0)$ is a strong hyper Markov law for the logistic model.

To begin this construction we require a strong hyper Markov law for the saturated model. One possibility is by extending (11) to larger 2-way tables.

THEOREM 4. *If a law $\pounds(\tilde{\theta})$ for a 2-way contingency table $X \times Y$ on $\mathcal{X} \times \mathcal{Y}$ has a density of the form:*

$$h\left\{\left(\frac{\theta_{xy}\theta_{x^*y^*}}{\theta_{xy^*}\theta_{x^*y}}\right)_{x\neq x^*, y\neq y^*}\right\}\prod_{x,y}\theta_{xy}^{\alpha_{xy}-1}, \tag{12}$$

for some $x^*, y^* \in \mathcal{X}, \mathcal{Y}$ and a positive Lebesgue intregrable function h, then it is strong hyper Markov.

Proof. Define $\theta_{+y} = p(Y = y \mid \theta)$ and $\theta_{x|y} = p(X = x \mid Y = y, \theta)$. Then the Jacobian determinant of the transformation $\theta_{xy} \mapsto (\theta_{+y}, \theta_{x|y})$ is

$$\left|\frac{\mathrm{d}\theta_{xy}}{\mathrm{d}(\theta_{+y}, \theta_{x|y})}\right| = \prod_y \theta_{+y}^{|\mathcal{X}|-1},$$

which gives the joint density for $(\theta_{+y}, \theta_{x|y})$:

$$\prod_y \theta_{+y}^{\alpha_{+y}-1}\, h\left\{\left(\frac{\theta_{x|y}\theta_{x^*|y^*}}{\theta_{x|y^*}\theta_{x^*|y}}\right)_{x\neq x^*, y\neq y^*}\right\} \prod_{x,y} \theta_{x|y}^{\alpha_{xy}-1}.$$

This factorizes into a term involving only θ_{+y} terms, and another involving only $\theta_{x|y}$ terms, and therefore $\tilde{\theta}_Y \perp\!\!\!\perp \tilde{\theta}_{X|Y}$. By symmetry, the same argument holds in the other direction. □

Theorem 4 can be viewed as the Bayesian counterpart to the theorem of Altham (1970), that the cross-ratio of a 2-way contingency table is variation independent of the marginal distributions.

It is unclear if the converse is true, *i.e.* if (12) characterizes all possible strong hyper Markov laws with full support. The corresponding result for (11) relies on results from functional equations, and these arguments can not be easily extended directly to higher dimensions.

Applying the conditioning approach to this law leads to the following law for the logistic model.

Example 3. We know from Theorem 4 that densities of the form

$$h\left\{\left(\frac{\theta_{x1}\theta_{x^*0}}{\theta_{x0}\theta_{x^*1}}\right)_{x\neq x^*}\right\} \prod_{x\in\mathcal{X}} \theta_{x0}^{a_{x0}-1}\theta_{x1}^{a_{x1}-1},$$

for some arbitrary $x^* \in \mathcal{X}$, are strong hyper Markov for the full $|\mathcal{X}| \times 2$ contingency table model.

The Jacobian determinant of the above transformation is

$$\left|\frac{\mathrm{d}\theta_{Y|X}}{\mathrm{d}(\alpha, \beta, \eta)}\right| \propto \prod_{x\in\mathcal{X}} \frac{e^{\alpha+\beta^\top x+\eta_x}}{(1+e^{\alpha+\beta^\top x+\eta_x})^2},$$

and hence the density for $\mathcal{L}'(\tilde{\alpha}, \tilde{\beta}, \tilde{\eta})$ is of the form

$$h\left\{\left(e^{\beta^\top(x-x^*)+\eta_x-\eta_{x^*}}\right)_{x\neq x^*}\right\} \prod_{x\in\mathcal{X}} \frac{e^{(\alpha+\beta^\top x+\eta_x)a_{x1}}}{(1+e^{\alpha+\beta^\top x+\eta_x})^{a_{x+}}}.$$

By conditioning on $\eta_x = 0$ for all $x \in \mathcal{X}$, we obtain the density of $\mathcal{L}(\tilde{\alpha}, \tilde{\beta})$:

$$g(\beta) \prod_{x\in\mathcal{X}} \frac{e^{(\alpha+\beta^\top x)a_{x1}}}{(1+e^{\alpha+\beta^\top x})^{a_{x+}}}, \tag{13}$$

where $g(\beta) = h\{(e^{\beta^\top(x-x^*)})_{x\neq x^*}\}$.

The Jacobian of the transformation in terms of the retrospective parameters is

$$\left|\frac{\mathrm{d}(\alpha, \beta, \theta_X)}{\mathrm{d}(\theta_{X|0}, \beta, \gamma)}\right| = \frac{(1-\gamma)^{|\mathcal{X}|-1}}{\gamma} \prod_{x\in\mathcal{X}} (1+e^{\alpha+\beta^\top x}),$$

and so the density of $\mathcal{L}(\tilde\theta_{X|0}, \tilde\beta)$ is

$$g(\beta) \frac{\prod_{x\in\mathcal{X}} \theta_{x|0}^{a_{x+}-1} e^{a_{x1}\beta^\top x}}{\left(\sum_{x\in\mathcal{X}} e^{\beta^\top x}\theta_{x|0}\right)^{a+1}}. \tag{14}$$

There are other ways to perform such a conditioning operation, such as using the odds ratio, but η has the desirable property of being invariant to the choice of x^* and $x_1, \ldots, x_{k+1}$.

The prior from Staicu (2010, Example 2) is obtained on rewriting (13) as

$$g^*(\beta) e^{\alpha a_{+1}} \prod_{x\in\mathcal{X}} \left(1 + e^{\alpha+\beta^\top x}\right)^{-a_{x+}},$$

where $g^*(\beta) = g(\beta) \exp\left(\sum_{x\in\mathcal{X}} a_{x1}\beta^\top x\right)$. On taking the limit as $a_{+1} \to 0$ we obtain the improper prior of Seaman & Richardson (2004) and Staicu (2010, Example 1).

However, we argue that the form of (13) is more easily interpreted: it can be thought of as the product of an improper prior with density $g(\beta)\, d\beta\, d\alpha$ and a logistic likelihood function, where the a_{xy} represent pseudo-counts. This has the further benefit of being able easily to adapt existing computational methods: for example, a Laplace approximation can be found using regular logistic regression software.

Although x appears in the density of $\mathcal{L}(\tilde\alpha, \tilde\beta)$, we disagree with Staicu (2010) that this constitutes a covariate dependent prior, such as the g-priors of Zellner (1986): it is dependent on the *a priori* expected frequency of the covariates, and not the observed frequency of the covariates in the data.

This law can itself be constructed as the posterior of a beta prior law.

PROPOSITION 1. *For each $x \in \mathcal{X}$, let*

$$\tau_x = \frac{e^{\alpha+\beta^\top x_i}}{1 + e^{\alpha+\beta^\top x_i}}.$$

For some $x_1, \ldots, x_{k+1} \in \mathcal{X}$ such that $(1, x_1), (1, x_2), \ldots, (1, x_{k+1})$ are linearly independent, let $\mathcal{L}'(\tilde\theta)$ be the product law of the marginal laws

$$\mathcal{L}'(\tilde\tau_{x_i}) = \mathcal{B}(a_{x_i 0}, a_{x_i 1}).$$

For all other $x \neq x_1, \ldots, x_{k+1}$, let

$$\mathcal{L}'(Z_x \mid \tilde\theta) = \text{Binomial}(a_{x+}, \tau_x).$$

Then the posterior law $\mathcal{L}'(\tilde\theta \mid Z_x = a_{x1})$ will have density of the form (13)*, where g constant.*

Proof. The prior law $\mathcal{L}'(\tilde\alpha, \tilde\beta)$ will have density proportional to

$$\prod_{x=x_1,\ldots,x_k} \frac{e^{(\alpha+\beta^\top x)a_{x1}}}{(1 + e^{\alpha+\beta^\top x})^{a_{x+}}}.$$

Likewise the likelihood of $(Z_x = a_{x1})_{x\neq x_1,\ldots,x_{k+1}}$ will be proportional to

$$\prod_{x\neq x_1,\ldots,x_{k+1}} \frac{e^{(\alpha+\beta^\top x)a_{x1}}}{(1 + e^{\alpha+\beta^\top x})^{a_{x+}}}.$$

□

This is particularly useful for implementing such procedures in generic Bayesian MCMC packages such as WinBUGS, OpenBUGS and JAGS: note that these packages happily accept

non-integer values for binomial counts. Furthermore, arbitrary functions g can be included by use of the "zero Poisson" trick: see Spiegelhalter et al. (2003, "Specifying a new sampling distribution").

Unfortunately, this method is somewhat impractical for large numbers of covariates. In particular, we note that the size of $\mathcal{X}$ increases exponentially with its dimensionality k. Furthermore, as $\mathcal{X}$ increases, $\tilde{\beta}$ will tend to concentrate around 0. To compensate for this, the values of (a_{xy}) can be chosen closer to 0, but unfortunately, the above software packages tend not work well, if at all, for very small values.

5. STRATIFIED CASE-CONTROL STUDIES

A more complicated case is that of stratified or matched case-control studies, in which participants are selected by both the outcome Y and an additional stratum variable S. Such a design can often estimate the odds-ratio of interest with much greater efficiency than an unstratified study.

The model is similar to that above, but with an intercept parameter that varies by stratum, so that the prospective model is

$$p(y \mid x, s, \alpha, \beta) = \frac{e^{\alpha_s + \beta^\top x}}{1 + e^{\alpha_s + \beta^\top x}}.$$

Unfortunately, this additional complication makes estimation more difficult. As the number of strata will increase with the sample size n, the usual maximum likelihood estimator is no longer consistent.

Instead, the standard classical approach seeks to maximize the conditional likelihood

$$\ell_{\mathrm{c}}(\beta) = \prod_{s \in \mathcal{S}} \frac{\prod_{i \in I_s} e^{y_i \beta^\top x_x}}{\sum_\rho \prod_{i \in I_s} e^{y_{\rho(i)} \beta^\top x_x}},$$

where $I_s = \{i : s_i = s\}$, and the summation in the denominator is over the possible permutations of $(y_i)_{i \in I_s}$.

If there are a cases and b controls in each stratum, called a:b matching, the sum in the denominator will have $\binom{a+b}{a}$ terms. In order to keep this computationally tractable, most studies use 1:1 or 1:m matching.

However for a Bayesian analysis the conditional likelihood does not have a direct interpretation. Rice (2004, Theorem 1) showed there exists a law such that the marginal retrospective likelihood $\bar{p}(x \mid y, s, \beta)$ is proportional to the conditional likelihood. However such a law depends on the matching scheme: *e.g.* a 1:1 matched design will require a different law than a 1:2 matched design.

Instead, we extend Theorem 2 to find conditions under which we can use the prospective likelihood for any matching scheme.

THEOREM 5. *Let $\pounds(\tilde{\theta}_{XY|S})$ be a prior law for the parameters of the stratified logistic model. Then the posterior marginal law for $\tilde{\beta}$ is the same under both the prospective and the retrospective likelihood, for all possible observations $(x^{(n)}, y^{(n)}, s^{(n)})$, if and only if*

$$\tilde{\beta} \perp\!\!\!\perp \tilde{\theta}_{X|S} \qquad \textit{and} \qquad \tilde{\beta} \perp\!\!\!\perp \tilde{\theta}_{Y|S} \qquad [\pounds].$$

Proof. The argument is essentially the same as that of Theorem 2, noting that $\theta_{X|S}$ and $\theta_{Y|S}$ are the joint distributions for the random vectors $(X|S = s)_{s \in \mathcal{S}}$ and $(Y|S = s)_{s \in \mathcal{S}}$, respectively. $\square$

To construct such laws, we use a conditioning procedure similar to that in the previous section. First, for each stratum s, let $\pounds_s(\tilde{\theta}_{XY|S=s})$ be a law satisfying Theorem 2, where $\theta_{Y|X,S=s} \simeq (\alpha_s, \beta_s)$. Then define $\pounds^*(\tilde{\theta}_{XY|S})$ to be the product law $\prod_s \pounds_s$, and therefore

$$\tilde{\theta}_{X|S} \perp\!\!\!\perp (\tilde{\beta}_s)_{s\in\mathcal{S}} \qquad \text{and} \qquad \tilde{\theta}_{Y|S} \perp\!\!\!\perp (\tilde{\beta}_s)_{s\in\mathcal{S}} \qquad [\pounds^*].$$

This can be reparametrized in terms of $\left[\beta, (\tau_s)_{s\neq s^*}\right] \simeq (\beta_s)_{s\in\mathcal{S}}$, where $\beta = \beta_{s^*}$ for some stratum s^*, and $\tau_s = \beta_s - \beta$ for each $s \neq s^*$. Finally, we condition on $\tau_s = 0$. Since

$$\tilde{\theta}_{X|S} \perp\!\!\!\perp \tilde{\beta} \;\Big|\; (\tilde{\tau}_s)_{s\neq s^*} \qquad \text{and} \qquad \tilde{\theta}_{Y|S} \perp\!\!\!\perp \tilde{\beta} \;\Big|\; (\tilde{\tau}_s)_{s\neq s^*} \qquad [\pounds^*],$$

it follows that $\pounds(\tilde{\theta}_{XY|S})$ defined as $\pounds^*(\tilde{\theta}_{XY|S} \mid \tilde{\tau} = 0)$ will satisfy the conditions of Theorem 5.

Example 4. If we let each $\pounds_s(\tilde{\alpha}_s, \tilde{\beta}_s)$ be of the form in Example 3, the density for the law $\pounds^*(\tilde{\alpha}, \tilde{\beta}, \tilde{\tau})$ will be of the form

$$\prod_{s\in\mathcal{S}} g_s(\beta + \tau_S) \prod_{x\in\mathcal{X}} \frac{e^{(\alpha_s + (\beta+\tau_s)^\top x) a_{x1s}}}{(1 + e^{\alpha_s + (\beta-\tau_s)^\top x})^{a_{x+s}}}.$$

Conditioning on $\tilde{\tau} = 0$ gives a density for $\pounds(\tilde{\alpha}, \tilde{\beta})$ as

$$g(\beta) \prod_{(x,s)\in\mathcal{X}\times\mathcal{S}} \frac{e^{(\alpha_s + \beta^\top x) a_{x1s}}}{(1 + e^{\alpha_s + \beta^\top x})^{a_{x+s}}}.$$

This is of the same form as the density (13), where the strata are treated as an additional categorical covariate in the model. Furthermore, the marginal laws $\pounds(\alpha_s, \beta)$ will also be of this form, and the stratum parameters $(\alpha_s)_{s\in\mathcal{S}}$ will be conditionally independent given β. Moreover, if the parameters are the same across strata (*i.e.* $a_{xys} = a_{xys'}$), then these stratum parameters are exchangeable, which could be a reasonable assumption in many analyses.

We have not specified a model for the stratum variable S, as we have assumed all data are observed conditional on S. However, under the additional assumption

$$\tilde{\theta}_{XY|S} \perp\!\!\!\perp \tilde{\theta}_S \quad [\pounds],$$

the data can be treated as if they were randomly sampled from the population, as would hold for a cross-sectional study.

6. DISCUSSION

A natural question is how to extend the above laws to the case where $\mathcal{X}$ is infinite, for example where a covariate is continuous. One obvious choice would be to replace the Dirichlet law for $\pounds(\tilde{\theta}_X)$ with a Dirichlet process. However the resulting density for $\pounds(\tilde{\theta}_{Y|X})$ in (13) would involve an infinite product, making it difficult to apply the standard Dirichlet process machinery of taking projections onto finite partitions of $\mathcal{X}$, and appealing to the Kolmogorov extension theorem.

There is potential for these techniques to be successfully applied to other models. In particular, the stratified case-control model is closely related to the Rasch model, commonly used in psychometrics for measuring ability or attitudes of individuals based on tests and questionnaires.

REFERENCES

ALDOUS, D. J. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, vol. 1117 of *Lecture Notes in Mathematics*. Berlin: Springer, pp. 1–198.

ALTHAM, P. M. E. (1969). Exact Bayesian analysis of a 2×2 contingency table, and Fisher's "exact" significance test. *J. R. Statist. Soc. B* **31**, 261–269.

ALTHAM, P. M. E. (1970). The measurement of association of rows and columns for an $r \times s$ contingency table. *J. R. Statist. Soc. B* **32**, 63–73.

ASHBY, D., HUTTON, J. L. & MCGEE, M. A. (1993). Simple Bayesian analyses for case-control studies in cancer epidemiology. *J. R. Statist. Soc. D* **42**, 385–397.

DAWID, A. P. & LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21**, 1272–1317.

DAWID, A. P. & LAURITZEN, S. L. (2001). Compatible prior distributions. In *Bayesian Methods with Applications to Science, Policy and Official Statistics*, E. I. George, ed. Office for Official Publications of the European Communities, pp. 109–118.

DAWID, A. P., STONE, M. & ZIDEK, J. V. (1973). Marginalization Paradoxes in Bayesian and Structural Inference. *J. R. Statist. Soc. B* **35**, 189–233.

GEIGER, D. & HECKERMAN, D. (1997). A characterization of the Dirichlet distribution through global and local parameter independence. *Ann. Statist.* **25**, 1344–1369.

GUSTAFSON, P., LE, N. D. & VALLÉE, M. (2002). A bayesian approach to case-control studies with errors in covariables. *Biostatistics* **3**, 229–243.

MARSHALL, R. J. (1988). Bayesian analysis of case-control studies. *Statist. Med.* **7**, 1223–1230.

MUKHERJEE, B., SINHA, S. & GHOSH, M. (2005). Bayesian analysis of case-control studies. In *Bayesian thinking: modeling and computation*, D. K. Dey & C. R. Rao, eds., vol. 25 of *Handbook of Statistics*. Amsterdam: Elsevier/North-Holland, pp. 793–819.

MÜLLER, P. & ROEDER, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika* **84**, 523–537.

NURMINEN, M. & MUTANEN, P. (1987). Exact Bayesian analysis of two proportions. *Scand. J. Statist.* **14**, 67–77.

PARK, M. Y. & HASTIE, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**, 30–50.

PATEFIELD, W. M. (1985). Information from the maximized likelihood function. *Biometrika* **72**, 664–668.

PRENTICE, R. L. & PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.

RICE, K. M. (2004). Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications. *J. Am. Statist. Assoc.* **99**, 510–522.

SEAMAN, S. R. & RICHARDSON, S. (2001). Bayesian analysis of case-control studies with categorical covariates. *Biometrika* **88**, 1073–1088.

SEAMAN, S. R. & RICHARDSON, S. (2004). Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Biometrika* **91**, 15–25.

SPIEGELHALTER, D., THOMAS, A., BEST, N. & LUNN, D. (2003). *WinBUGS User Manual*. MRC Biostatistics Unit, Cambridge, U.K. Version 1.4.

STAICU, A.-M. (2010). On the equivalence of prospective and retrospective likelihood methods in case-control studies. *Biometrika* **97**, 990–996.

WU, T. T., CHEN, Y. F., HASTIE, T., SOBEL, E. & LANGE, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714–721.

ZELEN, M. & PARKER, R. A. (1986). Case-control studies and bayesian inference. *Statist. Med.* **5**, 261–269.

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian inference and decision techniques*, P. K. Goel & A. Zellner, eds., vol. 6 of *Studies in Bayesian Econometrics and Statistics*. Amsterdam: North-Holland, pp. 233–243.